Guidelines for Developing a Data Dictionary

Save to myBoK

This practice brief has been retired. It is made available for historical purposes only.

Information systems are only as good as their data. Without a mutually agreed-upon set of data elements with clearly defined names and definitions, the validity and reliability of the data contained in a system are suspect at best and must be discounted at worst. The data dictionary and its relationship with the metadata registry are the foundation of an information system and the central building block that supports communication across business processes.

Data Dictionary Defined

To advance work toward electronic health record (EHR) content, AHIMA formed an e-HIM® work group to educate members and the industry on the importance of standardizing data content and data definitions within provider organizations and the industry as a driver to quality of care and patient safety. The work group defined a data dictionary as a descriptive list of names (also called representations or displays), definitions, and attributes of data elements to be collected in an information system or database. The purpose of the data dictionary is to standardize definitions and ensure consistency of use.

Rationale for Data Dictionaries

Standardizing data enhances interoperability across systems. It also improves data validity and reliability within, across, and outside the enterprise. Communication is improved in clinical treatment, research, and business processes through a common understanding of terms. Standardization provides developers with a common road map to promote consistency across applications.

Lack of a sound data dictionary can cause problems within and across organizations. Organizations may call the same data element by different names or they may call different data elements by the same name across an enterprise. As a result, an organization may not collect all of the information it needs or it may be unable to combine or map data across systems because the definitions are not identical. A worse possibility is that an organization may combine data elements it believes to be equivalent and draws incorrect inferences from the invalid data. Multiple users entering data may have different definitions or perceptions of what goes into a data field, thereby confounding the data (e.g., are “reason for visit” and “chief complaint” the same or different?).

Large complex systems with multiple stakeholders (internal and external) often require use of multiple, differing data sets. Variances among the data sets that are not recognized across the system can affect the information flow as well as the workflow. Maintaining expansive, overlapping data sets is costly to the organization in time and money and affects the quality of care. The organization will not be positioned for harmonizing information at the regional or national level.

Guideline Development Process

The work group conducted a comparative study of data definitions at the field definition level in order to create guidelines for developing a data dictionary. The purpose of these guidelines is to assist in building data dictionaries at the organizational level, aid in the development of new and existing data content standards, and support national standards harmonization efforts.

Since it is too early to know the impact of the federal data standards harmonization project sponsored by the Office of the National Coordinator for Health Information Technology, the work group centered its work around data dictionary development, whether new or updated. It is not too early for organizations to clean up their own houses through alignment of data content; this optimizes internal understanding as well as prepares for further alignment with the federal effort.

Taking care to select data sets affecting all care settings, the work group chose the following 11 major industry standard data sets for comparison:

- ASTM International's E1384-02a Practice for Content and Structure of the Electronic Health Record Minimum Essential Data Set
- ASTM International's WK4363 Standard Specification for the Continuity of Care Record (CCR)
- Doctor's Office Quality Information Technology's Data Element Specification v.1.1.2
- Electronic Medical Summary project (Canada) Core Data Set
- International Organization for Standardization (ISO)/TS 18308 Health Informatics: Requirements for an Electronic Health Record Architecture
- Joint Commission on Accreditation of Healthcare Organizations Comprehensive Accreditation's Manual for Ambulatory Care: Information Management Standards 6.20, EP1
- Centers for Medicare and Medicaid Services' Minimum Data Set, Version 2.0, for Nursing Home Resident Assessment and Care Screening
- National Center for Vital and Health Statistics' Core Health Data Elements
- Centers for Medicare and Medicaid Services and the Joint Commission on National Hospital Quality Measures
- AHIMA's Personal Health Record Minimum Common Data Elements
- Health Level Seven's Clinical Document Architecture, release 2

The work group selected common data content standards to compare at the field level. The group agreed that a sample of 10 data elements would be selected from each of a variety of data category types (e.g., service instance, patient, observation, providers, orders, care, treatment plan, encounter, problems) for comparison across the selected data sets. Initially, the work group chose ASTM International's CCR as a base data set from which to select a representative sample of data elements. It quickly became apparent that ASTM International's E1384-02a Minimum Essential Data Set was far more developed in detail and inclusiveness. As a result, it became the base against which other data sets were compared.

Using the information gained from this comparative study along with their collective expertise, the work group created the following guidelines to assist the industry in the development of data dictionaries.

Guidelines for Developing a Data Dictionary

1. Design a plan for the development, implementation, and continuing maintenance of the data dictionary.

Preplanning is imperative. The development of a data dictionary is part of a larger process. An information model must first be developed to align the workflow with information flow. This includes deciding what data are required, how the data will be used, who will use the data, and how the data will flow internally and externally, including communications with other entities.

This should be a collaborative process, and stakeholders should be encouraged to resist the temptation to collect data simply because they can. In the ideal scenario, data are captured once for use by multiple users. The end result of this data mapping is the ability of multiple entities to mine the same data source. Each will know the exact nature of the data element each is accessing. The plan should also include:

- The type of media (paper, electronic, spreadsheet, relational database) in which the data dictionary will be developed and maintained. The media choice may depend on the complexity of the enterprise system and the availability of resources.
- Adequate funding and staffing with clearly defined roles and responsibilities for development and ongoing maintenance of the data dictionary. Databases are dynamic and can be affected by new business lines, changes in national standards, and clinical advancements.
- Provisions to ensure that all licensing agreements are in order.
- Ongoing education and training of all staff as appropriate to their use of data elements and their definitions.

2. Develop an enterprise data dictionary that integrates common data elements used across an enterprise.

One purpose of the data dictionary is to provide consistency and understanding of common data across applications. Preplanning is a must to accomplish this at an enterprise level. A process must be clearly defined and key stakeholders

identified. The process requires collecting information or metadata (data about the data) on each data element found to be common across domains. It is important to define up front what needs to be done before starting the dictionary. This includes defining what metadata will be collected on each element as well as what will not be collected. Examples of metadata include name of element, definition, application in which the data element is found, locator key, ownership, entity relationships, date first entered system, date element terminated from system, and system of origin.

A metadata registry is an authoritative source of reference information about the representation, meaning, and format of data collected and managed by an enterprise. It does not contain the data itself but the information that is necessary to clearly describe, inventory, analyze, and classify data.

3. Ensure collaborative involvement and buy-in of all key stakeholders when data requirements are being defined for an information system.

Stakeholders include data creators, data owners, and data users, both internal and external to the organization. Representation should reflect all geographies (departments, facilities, satellites, corporate representatives, and external entities). Each organization must identify its stakeholders based on its own unique business model, organizational structure, information flow, and reporting requirements. Different stakeholders may have different data element definitions within their local domain. Every attempt should be made to promote collaborative agreement whereby a datum is collected only once even though it may be used by multiple end users.

Take for example a large enterprise that discovered it had approximately 40 different representations for data elements with a set of values of “yes” and “no” throughout its data dictionary. These included: Y = yes, N = no; yes, no; 1 = yes, 0 = no; 1 = no, 0 = yes; 1 = yes, 2 = no. These should be standardized as one set of values in the enterprise data dictionary.

Public health and research are examples of external stakeholders. Public health reporting is often forgotten in the data requirements definition phase. As a result, organizations incur extra costs to develop special interfaces and maintain crosswalk tables to meet public health requirements.

The collaboration of all data stakeholders (e.g., clinical specialties, support services, HIM services, IS services, reimbursement specialists, administrative, legal, and public health agencies) should enhance consensus and understanding of data and their flow across all domains.

4. Develop an approvals process and documentation trail for all initial data dictionary decisions and for ongoing updates and maintenance.

It is important to document decisions made about the data dictionary throughout the life of the system. Each subsystem (e.g., finance, lab, radiology) should have one authoritative owner responsible for tracking all implemented data dictionary activations, deactivations, relevant dates, events, and decisions.

There must be a maintenance and change control process for adding new values, elements, and enactment dates. The subsystem owner should review and approve any additions to the system and integrate those changes through a collaborative process with other owners into the whole enterprise system. The process should address how a new datum applies in the local setting or domain and across all aspects of the enterprise.

5. Identify and retain details of data versions across all applications and databases.

Ensure clear mapping instructions for organization-specific definitions. Version control is essential for maintaining data reliability. It is important that the data set version is clearly identified. Differences between versions may be minor or extensive. It is critical that everyone in the enterprise operate on the same version in order to maintain data integrity and continuity. Version control is essential for data dissemination in standard format to satellite or remote facilities. Separate tables may be considered for keeping track of changes such as additions, deletions, and their relative effective dates.

6. Design flexibility and growth capabilities into the data dictionary so that it will accommodate architecture changes resulting from clinical or technical advances or regulatory changes.

Build expansion capabilities into the fundamental design to accommodate a dynamic system. There should be a plan for future expansion, such as expanding a data field from one element to multiple elements. Expansion must be carefully addressed because of the potential ramifications of concept migration, the change of an idea or concept over time through growth or change to the system. This becomes problematic when comparing data across time if the meaning of a particular element has changed while its name or representation has not. If a data element is totally revamped, document when that specific data element went into effect and when it was deactivated. If the data element expands into something new, do not migrate the old concept but rather create a new element to move forward. This will affect how the data are stored and retrieved. It may require consultation with vendors where current system limitations exist.

Always strive for concept permanence. Never reuse a concept even if it becomes obsolete. For example, when an ICD code number is retired, never reassign the retired code to a new concept. Always follow the defined coding practices. This becomes particularly important in data comparison. Address architecture flexibility in vendor contracts to allow for system upgrades and room for expansion to accommodate requirements common to provider-specific issues, user groups (multiple sites), or state-based directives.

7. Design room for expansion of field values over time.

Consider future needs to collapse and expand values to accommodate mapping from a larger to smaller or smaller to larger number of values within a field definition. When setting up the information system, consider how to accommodate multiple systems and how to go from one code system to another. Mapping and transferring guidelines should be clarified between data sets. For example, race or ethnicity is frequently defined with different values. One data set has four items, another has six. The mainframe or core system needs the maximum amount of values. The mapper needs to know the rules to use when collapsing six values into four. Migrating four to six is usually impossible, which creates other issues.

Gender is another core data element that can generate much discussion. Many systems only allow for male and female, while others provide for unknown and other. When an “other” category is an option, there should be a process for monitoring what is captured under that heading. When large numbers begin to appear in the category, there should be a review to determine if a new discrete category is required or if there is misunderstanding in the definition of the core element.

Take for example a data dictionary that must accommodate the changes necessary to adopt the current ICD-9-CM diagnosis code fields from six characters to what will be required for ICD-10-CM. Some organizations have been proactive and already made these changes and updated their data dictionary.

8. Follow established ISO/International Electrotechnical Commission (IEC) 11179 guidelines or rules for metadata registry (data dictionary) construction to promote interoperability and automated data sharing.

Uniformity of approach in data dictionary development avoids industry fragmentation. In an effort to promote and improve international communications among governments, businesses, and scientific communities, ISO and IEC have developed standards for specification and standardization of data elements. The ISO/IEC 11179 standard consists of:

- A framework for the generation and standardization of data elements
- A classification of concepts for the identification of domains
- Basic attributes of data elements
- Rules and guidelines for the formulation of data definitions
- Naming and identification principles for data elements
- Registration of data elements

This standard provides excellent detailed information and examples of how to classify and define data elements. It also includes examples of pitfalls and practices to avoid.

9. Adopt nationally recognized standards and normalize field definitions across data sets to accommodate multiple end user needs.

It is important to define all data characteristics to be included for each data element for all domains. This includes specifying domain boundaries and identifying linkages across domains. This will require extensive discussion and agreement among all stakeholders. The ideal is the development of a common integrated data and terminology model. Terminologies should be

coordinated to eliminate overlaps, redundancies, and inconsistencies. This will eliminate the need for mapping among terminologies.

10. Beware of differing standards for the same clinical or business concepts.

Do not assume that things labeled the same are actually identical or will map one to one. For example, there are several different wound staging protocols. The Centers for Medicare and Medicaid Services require one version in the Minimum Data Set Version 2.0 for reimbursement purposes. For clinical care, it requires a different staging protocol that is based on the AHRQ Clinical Practice Guideline for Pressure Ulcers. MDS 3.0, currently in beta with an expected release date in 2007, is expected to remedy this particular problem by requiring only one standard. Pain measurement scales are another example of multiple scales for the same concept. Always check with a subject matter expert to ensure valid data.

11. Use geographic codes and geocoding standards that conform to those established by the National Spatial Data Infrastructure and the Federal Geographic Data Committee, following the guidelines of the Federal Information Processing Standards.

Valid street addresses, zip codes, county, state, and country codes are important to information exchange across systems and geopolitical boundaries. Standardization of geographic codes enhances interoperability of systems. Healthcare uses this information for tracking diseases as well as people. Using internationally accepted standards further enhances the interoperability of systems and the exchange of information. The following are recommended resources for geographic codes:

- Federal Information Processing Standards (www.itl.nist.gov/fipspubs)
- Federal Geographic Data Committee (www.fgdc.gov)
- United States Postal Service (www.usps.gov)
- National Spatial Data Infrastructure (www.fgdc.gov/nsdi/nsdi.html)
- International Organization for Standardization (www.iso.org)

12. Test the information system to demonstrate conformance to standards as defined in the data dictionary.

Once the data dictionary is completed, a test plan should be developed to ensure that the system implementation supports the data dictionary. This should include sampling data inputs and outputs for conformance, validity, and reliability. This process should also verify interoperability of systems.

13. Provide ongoing education and training for all staff as appropriate to their use of data elements and their definitions.

To ensure consistency of understanding, application, and use of data, it is imperative to provide ongoing education in those definitions. New employee orientation should routinely include exposure to the concepts expressed in the data dictionary.

14. Assess the extent to which the use of the agreed-upon data elements supply consistency of information sharing and avoid duplication.

Ensure simultaneous adoption of new knowledge developed through research and changing terminologies reflective of changes in clinical practice. Specific stakeholders external to most end-user organizations that should be involved in the development and modification of data elements that affect clinical care include all American Board of Medical Specialty recognized specialty societies (e.g., American Academy of Pediatrics and the American Academy of Family Physicians). This evaluation and modification process should be ongoing and involve members of the specialty societies at all stages of the process.

Conclusion

The creation and maintenance of the data dictionary is pivotal to the success of an EHR system. Much thought and effort must go into the planning and the maintenance of this foundational information. Collaboration and buy-in by stakeholders across all domains is critical to the success of the EHR implementation. A process for ongoing maintenance and updates as well as version control must be in place. The upfront design must provide room for change, growth, and expansion over time. Organizations should follow established guidelines such as the ISO/IEC 11179 and the geographic code systems where possible to promote interoperability. Normalization of concepts across end users is an ultimate goal, while any variances in

business or clinical concepts should be carefully noted. Once the hard work of the build has been completed, the EHR system should be thoroughly tested to ensure it accurately reflects the standards as defined in the data dictionary.

Prepared by

AHIMA e-HIM Workgroup on EHR Data Content
Carol Adam
Dena Barley
Robert Bishop, MBA, PMP
Keith W. Boone
Christine Brooke, RHIA, CHP
Kathy Callan, MA, RHIA
Barbara Demster, MS, RHIA
Kathy Giannangelo, RHIA, CCS
Matthew Greene, RHIA, CCS
Beth Hjort, RHIA, CHPS
Laurie Peters, RHIT, CCS
Christine Rooker, MA, RHIA, CTR
Barbara Samuels, MBA, RHIA
Carol Schuster, RHIA, MSM
Mary H. Stanfill, RHIA, CCS, CCS-P
Dolores Stephens, MS, RHIT
Hao Wang, PhD, MPA
Elmer (Lee) Washington, MD, MPH
Lou Ann Wiedemann, MS, RHIA
Margaret Williams, AM
Carolyn Wilson, MBA, RHIA
Pat Wilson, RT(R), CPC

Article citation:
AHIMA e-HIM Work Group on EHR Data Content. "Guidelines for Developing a Data Dictionary" *Journal of AHIMA* 77, no.2 (February 2006): 64A-D.

Driving the Power of Knowledge

Copyright 2022 by The American Health Information Management Association. All Rights Reserved.